

# The Research Platform for Building Medical Diagnostic Services

Yury Apanasik, Irina Shabalina, Larisa Kuznetsova  
 Petrozavodsk State University (PetrSU)  
 Petrozavodsk, Russia  
 {apanasik, i\_shabalina, l\_kuznetsova}@petsu.ru

**Abstract**—This paper describes research and development of a software platform for processing medical diagnostic data. This work is carried out by the IT-park of PetrSU in collaboration with the Faculty of Medicine. The aim of the Platform development is to present universal software tools for research done on the basis of the initial information on the diseases, and the mathematical methods of data processing. The paper describes the software platform general architecture, as well as the methods of diagnostic rules creation using the iteration model and the poly-procedural approach. The offered methods have sufficient generality and can be applied in research of a wide range of diseases.

**Keywords**—*Mathematical methods of diagnostics, Statistical methods, Software design.*

## I. INTRODUCTION

Currently, a significant number of researches in the field of medicine are aimed at discovery of new medications, development of new diagnostic methods and disease treatment, discovery of cause and effect relations in disease progression [1]. During the follow-up of a patient, test results and information on the dynamics of disease progression are collected and analyzed. Diverse diagnostic information incoming from multidisciplinary doctors helps to create a full image of the medical condition of a patient. Diagnostic data arrays received in this manner can be used in research, for example in studying a specific disease. In addition to that, in order to solve the problems in the given application area, a practice of diagnostic data research is developed. The practice is based on both standard and special methods and approaches to data analysis [2]. Most of these “individual” practices can be extended to the form of universal practice of processing medical diagnostic data, and its methods can be implemented in the form of software services of a unified software research platform (hereafter – Platform).

The aim of the Platform development is to present universal software tools for research done on the basis of the initial information on the diseases, and the mathematical methods of data processing. The services of the Platform are aimed to solve the following main problems:

- Initial collection and systematization of medical survey data (test results, analyses, diagnoses, prescriptions, procedures, etc.);
- Creation of a database with build-in tools of collecting data accumulated not by just one researcher, but by some medical society, in order to solve the applied research problems;
- Solution of research problems, in particular developing diagnostic procedures on the basis of the mathematical methods of data processing.

The paper presents a project of research platform architecture, purpose of various architecture levels, and general requirements for their implementation. Besides, it presents the description of the functions, which the platform performs, and possible applications of the platform. Significant attention is paid to medical diagnostic services composition; the mathematical methods which allow creating diagnostic algorithms are presented.

## II. RESEARCH PLATFORM ARCHITECTURE

The main idea of the design of the research platform software architecture was to minimize the limitations of its technical use, as it makes maximally wide application possible. A multilayer architecture (see Fig. 1) is offered, which contains the following levels: “data storage”, “data access”, “data adaptation”, “applied services”.

The research platform is presented as both hardware for operating the medical data, and in the form of protocols and rules, which are implemented in the applied services that are the main tools of a medic researcher.

### A. Data storage level

This level is presented by a given set of data storages of strong typing. Support of the following minimal set of the types of storage is required for the platform work:

- 1) *Medical data storage*. This type of storage may be presented as a relational database containing medical data tagged to a specific patient (object). The format described by the open standard CCR (Continuity of

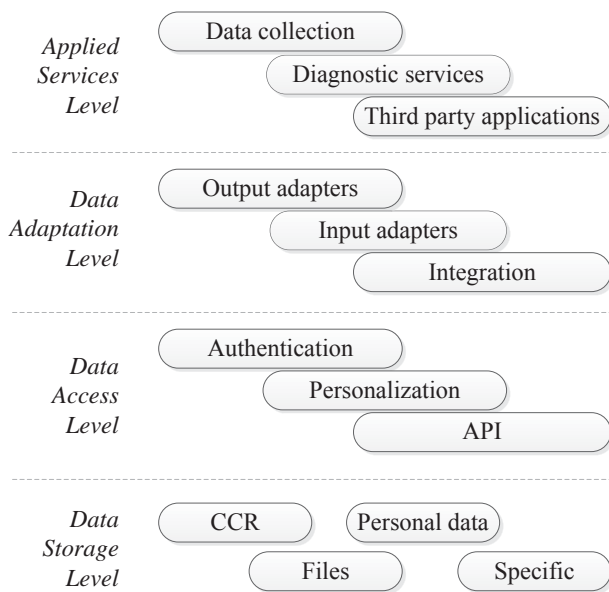


Fig. 1. Research platform architecture

Care Record), developed by a group of research centers in the USA [3] can be used as a format of storing. CCR is designed for storing private medical data, divided in such blocks as: Payers, Problems, Social History, Alerts, Medications, Immunizations, Vital Signs, Results, Procedures. CCR allows describing the medical data on the object to the full extent and is widely spread: such medical services as Google Health and Microsoft HealthVault are based on it. As alternative the Clinical Document Architecture (CDA) of Health Level 7 (HL7) standard can be supported by data storage.

- 2) *Personal data storage.* As the research platform does not have any accounting functions, it is more convenient and correct to store all the medical data in anonymized form. However with the use of this approach it is impossible to compare the data in dynamics, so a hybrid approach is offered, when the medical data is stored tagged to anonymized codes of the patients, and the data on the tag codes to specific people are stored in separate storages.
- 3) *File storage* is designed for data storage in the form of files. As far as it is possible, using this type of storage should be restricted, as using non-typed data from the files is complicated while performing analytical tasks.
- 4) *Specific storage.* Storages aimed at storing highly specialized information, with their support done separately, belong to this category. The database of on-line monitoring system of a patient's medical condition can be an example of this type of storage. In particular the project CardiaCare is a system of such kind. In the course of this project a hardware

and software complex is developed for real-time monitoring of cardiac function [4].

*B. Data access level*

The level is presented by a software layer allowing access to the data in the storages of various types. Data access level should provide solution for the following problems:

- 1) With the use of the tools of this level the details of implementation of the data storage methods should be hidden: number and type of the storages, physical addresses of the storages. The levels located higher in the architecture hierarchy should gain access to the data not directly but using the tools of this level.
- 2) At the data access level problems of authentication and personalization should be solved in order to separate the stored data according to their belonging to different research groups. Permission to use own research data should be presented clearly, and until this moment all the information should be available only to its owner.
- 3) At the data access level information on all the connected to the platform databases should be included. At this level the tools of connecting additional storages and controlling the existing should be implanted.
- 4) Upon the delivery of new medical data, software of the data access level should divide these data between the proper storages, tag the data to the anonymous codes, using the data from the personal data storage.

*C. Data adaptation level*

The data in the way they are stored at the data storage level are not always easy for analyzing and using in various medical services. For this reason a layer of data adaption is separated in the research platform architecture, and is presented by a set of software adapter agents. Software agents perform selection of the data and transform them into one or the other format. And conversely they get the data from the external environment and save them in the storage in the appropriate format.

A software agent that performs data selection in order to find relationships between the symptoms in specific disease diagnosis can be mentioned as an example of an outswapping adapter. This software agent chooses all the entries on the objects with the given diagnosis and forms a limited set of concomitant symptoms that later can be analyzed by the diagnostic systems.

The main function of the data adaptation level is integrating the research platform database into the exterior medical software systems, for example, clinical information systems (CIS) of healthcare institutions. The integration can be one-sided or dual sided, and performed

by the means of implementation of a special software adapter. This adapter should support formats of data structuring both of the research platform and the external software environment.

*D. Applied services level*

Applied services are a set of tools which can be used by a medic researcher in his or her work. These tools include analytical systems, data collection services, as well as software products which supplement or use the information from the platform database in their work.

In the course of the research the main types of applied services are the software for statistical analysis of examination data with the purpose of discovering the relationships between the symptoms and finding the regularities. A separate class of such systems is the diagnostic services that form, on the basis of found regularities, diagnostic rules, which allow running diagnostics for the set of symptoms and getting the evaluation of the probabilities of potential diagnoses in computer-assisted way.

The basic service of the applied level is a program module which implements the model of establishing diagnosis according to the given set of symptoms, which is widely spread in medicine. Previously, the research of connective tissue dysplasia in children (hypermobility syndrome, HS) [5] was used as a subject field for approbation of such model. Previously offered mathematical models of diagnosis [6] including those developed on the basis of the Basic approach [7], served as a basis for the software of the Platform diagnostic services composition.

III. MEDICAL DIAGNOSTIC SERVICES BUILDING SYSTEM

The composition system of the medical diagnostic services is an important applied tool in the structure of the medical research. The system allows recognizing statistical relationships between the symptoms and diagnoses, and generating a software service that allows running computer-assisted diagnosis on the basis of the data on the disease collected by a researcher.

The system architecture (see Fig. 2) was built in accordance with the rules of the medical research platform. The data that are operated by the applied part of the system are outswapped to a separate database with the use of specifically developed software adapter. The adapter in its turn uses the software tools of the platform to extract the data from the storage. For the work of the diagnostic methods described in the part C of the paper, the data are outswapped by the adapter in a special way: for the given disease all the examination entries are extracted under the specified set of criteria (symptoms) with known diagnosis, and checked and approved by an expert (researcher).

Outswapped data form a teaching selection for the mathematical methods of diagnosis.

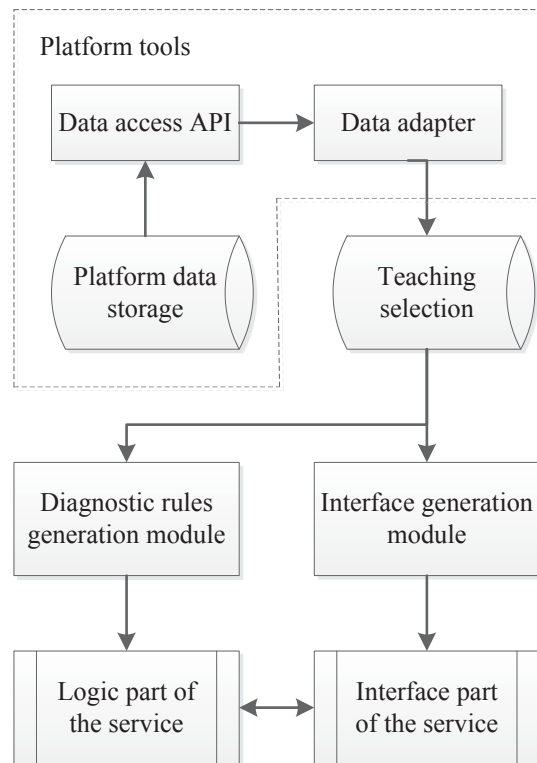


Fig.2. Architecture of the system of the diagnostic services composition

In the structure of the software component of the system of the medical diagnostic services composition two big modules can be separated: “diagnostic rules generation” and “interface generation”.

Diagnostic rules generation module uses build-in mathematical methods for creation a software library with clearly defined exterior interfaces permitting to get the evaluation of probabilities of the diagnosis for the given patient’s examination data. Obtained library performs its functions relating to the disease and its research.

Interface generation module creates a template interface for the user of the diagnostic service using the given in the teaching selection of the symptoms and diagnoses. Tools of the interface permit to input the examination data and get the evaluation of the probabilities of diagnoses using the tools of the diagnostic library. The module generates interface component for the Web as it is best suited for providing and distributing the diagnostic service. The example of interface part of diagnosis service for hypermobility syndrome is presented on Fig. 3.

Division of the application part of the system into two compounds allows, if necessary, develop a more complicated, different from the template, service interface based on the same logic.

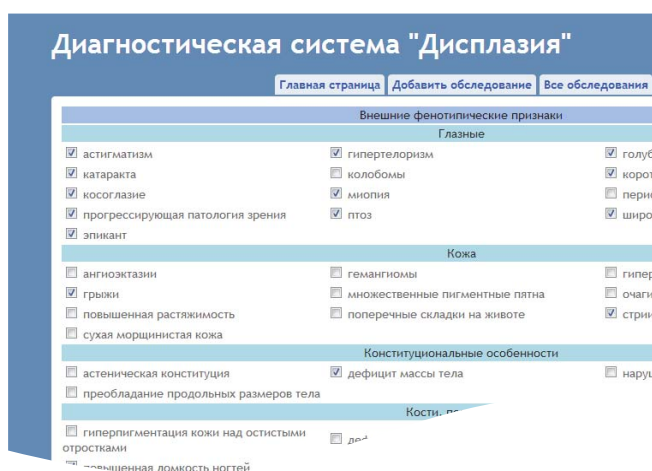


Fig. 3. Interface of diagnosis service for hypermobility syndrome

The system of the medical diagnostic services composition makes it possible for a medic researcher to summarize the results of the research in the form of completed software service without involvement of IT specialists. This system can be implemented as a decision support system in medical practice.

#### IV. MATHEMATICAL MODELS IN MEDICAL DIAGNOSTICS

##### A. Setting up a problem of medical diagnostics

In order to describe the mathematical methods of diagnostics it is necessary to perform formalization of the object region and develop a mathematical model of the problem. All the subject area objects (patients) comprise a set  $\Omega$  – that is a general set of objects. Part of the objects that have been diagnosed by experts comprise a teaching selection  $\Omega^* \subset \Omega$ . Each object  $X_i$  from the general set  $\Omega$  is a  $m$ -dimensional vector  $(x_{1i}, \dots, x_{mi})$  of test results based on the set of symptoms  $M$  ( $m = |M|$ ).

Vector column that contains test result data based on the set of symptoms is called **medical record** of the patient (object). Test result  $x_{ij}$  can be represented as:

$$x_{ij} = \begin{cases} x_{ij}^* \in Q_{ij}, \text{ object } i \text{ was} \\ \text{tested for symptom } j \\ \emptyset, \text{ otherwise} \end{cases} \quad (1)$$

All teaching selection objects have been diagnosed by experts. Thus teaching selection  $\Omega^*$  is divided into  $s$  number of non-overlapping subsets  $A_s$  based on index set of possible diseases  $S$ . Set  $A_s$  ( $s \in S$ ) that contains objects with diseases  $s$ , is called « $s$  class». Let  $N = \sum_{s \in S} N_s$ , where  $N_s$  is the number of objects in teaching selection that have diseases,  $N$  is the total number of objects in teaching selection.

When examining hidden dependencies of initial data in the teaching selection  $\Omega^*$  it is necessary to build a **decision**

**rule  $\Phi$**  to precisely assign object  $X_0$  to one of the classes  $A_s$  ( $s \in S$ ).

$$\Phi(X_0) = s^* \mid X_0 \in A_{s^*} \quad (2)$$

Further on, two methods that form the basis of the diagnostic abilities of the system of the diagnostic services development are described. The first one, the iteration model – is a diagnostic conception responsible for cycling through a number of steps. The iteration model requires the solution of three subproblems, one of which is performing diagnostics under the given set of symptoms. At this step it is feasible to run diagnostics by different methods, and then draw a general conclusion based on the results of work of each of the methods. In order to output the most statistically valid diagnosis from the set of the results, a poly-procedural approach is offered for use [6].

##### B. Diagnostic iteration model

The model is a step-wise diagnostic algorithm with actions divided into two groups: actions done by a human being (diagnostic service user), actions done by computerized system. Iteration model can be technically presented in the form of the following algorithm [5]:

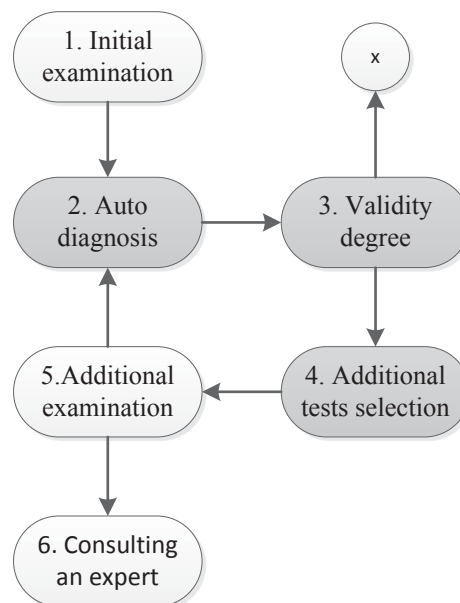


Fig. 4. Diagnostic iteration model

Step 1: Initial examination of object  $X^*$  in the test set  $T_0 \subset T$ , where  $T$  – is a set of all medical tests of the relevant disease.

Step 2: Computerized disease diagnostics based on a set of test results  $T_i$ , where  $i$  – is the number of diagnostic cycle iteration (starting from 1).

$$T_i = T_0 + \bigcup_{j=1}^i T_j^* \quad (3)$$

Diagnostics result is a vector of degree of membership  $\xi_k^i$  for each of the diagnostic classes  $A_k \in S$ .

Step 3: Evidence degree evaluation of the diagnosis  $\Theta_i(X_0)$  assigned at Step 2. If evidence degree is  $\Theta_i(X_0) = 1$  and the subject of checkup is not interested in further clarification of the diagnosis, it is the end of the algorithm; otherwise go to Step 4.

Step 4: Selection of additional tests based on a number of symptoms of set  $T_i^*$  implementation of which enables maximum clarification of the diagnosis. If the selection of additional tests is not possible, it is the end of the algorithm (the last diagnosis is considered to be certain, the system suggests consulting an expert); otherwise go to Step 5.

Step 5: Checking object  $X^*$  with a set of additional tests  $T_i^*$  and then going to Step 2.

It is required to solve 3 particular independent subproblems to implement iteration model:

- 1) Finding diagnostic method based on incomplete checkup data.
- 2) Finding evidence degree evaluation method for the diagnosis based on incomplete checkup data.
- 3) Finding a method for selecting additional tests, which satisfy external limits in the best way, maximizing the accuracy of the follow-up diagnosis.

### C. Diagnostic methods

The frequency methods, entropy method and likelihood ratio method were employed for diagnostics [6]. The usage of several methods increases the reliability of automate diagnostics.

The basis of frequency methods are Bayes decision rule theory [7, 8], the consideration of correlation between the properties was also implemented. This method uses the following decision rule:

$$\begin{aligned} \text{if } s^* &= \operatorname{argmax}_s \left\{ \frac{p(X \cdot A_s)}{\sum_{q \in S} p(X \cdot A_q)} \right\}, \\ \text{then } X &= (x_1, x_2, \dots, x_m) \in A_{s^*} \end{aligned} \quad (4)$$

where  $p(X \cdot A_s)$  is the probability estimate of occurrence in  $A_s$  of object  $X$  with symptoms  $x_1, x_2, \dots, x_m$ .

Entropy methods and likelihood ratio methods utilize the decision rule based on linear combination of symptoms:

$$\begin{aligned} \text{if } \sum_{i=1}^m \omega_i x_i &\in R_{s^*}, \\ \text{then } X &= (x_1, x_2, \dots, x_m) \in A_{s^*} \end{aligned} \quad (5)$$

where  $\omega_i$  is the weight of symptom ( $i = 1..m$ ),  $R_{s^*}$  is the range of linear combination values which match to  $s^*$  class.

The foundation of entropy methods is usage of the information measure [9]:

$$\omega_k = -(p_k \log_2 p_k + (1 - p_k) \log_2 (1 - p_k)) \quad (6)$$

where  $p_k$  is the probability of symptom  $k$  ( $k = 1..m$ ).

The likelihood ratio methods use following formula for weights  $\omega_k$  calculation:

$$\omega_k = \log_2 \frac{p(k / A_{s_1})}{p(k / A_{s_2})} \quad (7)$$

where  $p(k / A_{s_q})$  is the probability of symptom  $k$  for class  $A_{s_q}$  [10].

All of presented methods were implemented in medical diagnostic services building system described above.

### D. Poly-procedural approach

When coming to diagnostic conclusion the set  $R$  of diagnostic methods is used [5]. Each diagnostic method ( $r \in R$ ) establishes one diagnosis  $s$  out of the set of diagnoses  $S$  for the object. If using the method  $r$  classify all the objects included in the teaching selection, then  $N_{rs}$  of correctly recognized object are the objects for which the classification of the method corresponds to the classification of the expert. Other objects will be recognized by the method  $r$  incorrectly or fail to be recognized. Degree of reliability of the method  $r$  in recognition the object from the  $s$ -class is a statistical evaluation of probability of correct recognition of the object from  $A_s$  using the method  $r$ , which is calculated by the formula  $p_{rs} = N_{rs} / N_s$ . The probability of the method recognizing the object from  $A_s$  incorrectly or failing to recognize it can be measured by the value  $(1 - p_{rs})$ .

For establishing the most reliable diagnosis we will use the independence property of the employed diagnostic methods. The probability of the object  $A_s$  being correctly recognized using at least one of the set of methods  $R$  is calculated in the following way:

$$p_s = 1 - \prod_{r=1}^R (1 - p_{rs}) \quad (8)$$

It can be shown that  $p_s > p_{rs}$  for  $\forall r \in R$  and  $s \in S$ , which means that the reliability of diagnosis  $p_s$ , calculated by the formula (8) is higher than the reliabilities of diagnoses  $p_{rs}$ , calculated by separate methods.

The diagnostic results of the disease  $S$  while using methods for the object  $i$  ( $i \in N$ ) can be described by vector  $(z_{11}^i, z_{12}^i, \dots, z_{1s}^i, z_{21}^i, z_{22}^i, \dots, z_{2s}^i, \dots, z_{R1}^i, z_{R2}^i, \dots, z_{RS}^i)$ , the elements of which are calculated by the following way:

$$z_{rs}^i = \begin{cases} 1, & \text{if the object } i \\ & \text{is diagnosed } s \\ & \text{by method } r \\ 0, & \text{otherwise} \end{cases} \quad (9)$$

It should be mentioned that the method  $r$  calculates one diagnosis and in this case the following condition is fulfilled: if  $z_{rs}^i = 1$ , then  $z_{rs}^i = 0$ ,  $\dots$  O M O

Applying the formula (8) we can get the probability of the examined object  $i$  recognized as an object of the class #

$$L_{L s F} : s F L ; \quad (10)$$

@5

The value  $L$  represents the reliability of the poly-procedural diagnosis for the examined object.

The value  $V_L$  corresponds to the situation of incorrect recognition, which means that the diagnosis of the method differs from the diagnosis of the expert. In this case – the error probability is equal to 1, which is considered in the formula (10) in the following way:  $: s F L ; L : s F L ; L_s$

Among the evaluations  $L (s \in S)$ , obtained by the formula (10) the highest is separated, and corresponding to it value  $O$  should be considered the most reliable poly-procedural diagnosis for the object  $: E O L = N C \Delta =$  This approach to the calculation of the reliability of the diagnosis is called the poly-procedural approach [5].

*E. Results of methods implementation*

The combination of several diagnostics methods and poly-procedural approach were implemented and tested while researching connective tissue dysplasia syndrome (Hypermobility syndrome, HS).

Training sample contained 1000 objects and sample contained 200 objects were used for probation of methods. The set of 64 symptoms of HS was determined for each object of the samples. The methods have to determine three degrees of HS:  $O L s, t, u$

The poly-procedural approach requires preliminary usage of several diagnostics methods, so the frequency methods, entropy methods and likelihood ratio methods were employed for diagnostic [6].

Follow example presents the poly-procedural approach utilization. The results  $V$  of diagnostics methods usage for object are contained in Table I.

TABLE I. RESULTS OF DIAGNOSTIC METHODS

v	NL s	NL t	NL u	NL v
OL s	0	1	1	1
OL t	1	0	0	0
OL u	0	0	0	0

As shown in the table, the result of diagnostic by the method  $N L s$  disagree with others. The methods reliabilities  $L$  are presented in Table II.

TABLE II. METHODS RELIABILITIES

L	NL s	NL t	NL u	NL v
OL s	0,84	0,89	0,91	0,85
OL t	0,89	0,93	0,68	0,68
OL u	0,68	0,77	0,77	0,77

The Table III lists results of preliminary diagnostics by methods  $N L s, t, u$ , for the examined object. Results of calculation the reliability of the poly-procedural diagnosis  $L$ , experts diagnosis and poly-procedural diagnosis.

TABLE III. PART OF TABLE WITH POLY-PROCEDURAL DIAGNOSIS

E	Ns	Nt	Nu	Nv	$L_s$	$\frac{q}{6}$	$L_t$	O	O <sub>o</sub>
1	2	1	1	1	0,999	0,890	0,000	1	1
2	2	3	3	1	0,850	0,890	0,947	3	2
3	1	1	1	2	0,998	0,680	0,000	1	1
4	3	3	2	3	0,000	0,680	0,983	3	3

Recognition rate on a percentage for evaluation of the diagnostics quality was calculated after the methods employment to training and control samples. Test results are presented in Table IV.

TABLE IV. TESTS RESULTS

Method	Training sample	Test sample
Frequency method	83,3%	81,5%
Frequency method with correlations	88,5%	78,5%
Likelihood ratio method	77,6%	82%
Entropy method	76%	79,5%
Poly-procedural approach	86,4%	83,5%

The quality of suggested approaches diagnostics is about 88,5% correct recognized objects on the training sample and about 83,5% on the test sample. The poly-procedural approach consolidates diagnosis of several methods and increase the diagnostic reliability by 10-15% compared with each separate method.

V. CONCLUSION

The methods of diagnostic rules development and the poly-procedural approach have passed evaluation test while tissue dysplasia syndrome researching. The iteration model allows to conduct the diagnostics by each method and to make the selection of additional medical tests that are required for follow-up diagnosis.

During the research an architecture solution for the Platform has been offered, and the software modules based on the mathematical methods have been developed. These methods allow composing services for diagnostic

procedure for diseases, using the existing information on the examination results.

The Platform software architecture purposes to make maximally wide application utilization as a data bank for medical research. Software tools of Platform allow to make double-sided integration with clinical information systems and to receive data for different goal services development.

Such services are the decision support systems in the daily work of general practice doctors, and with appropriate adaptation they can be used for health condition self-testing by a wide range of users.

#### ACKNOWLEDGMENT

The article was published with financial support of the Strategic Development Program of Petrozavodsk State University.

#### REFERENCES

- [1] X.-H.Zhou, N. A. Obuchowski, *Statistical Methods in Diagnostic Medicine*.Wiley. 2011.
- [2] P. Armitage, *Statistical Methods in Medical Research*.Wiley. 2002.
- [3] Standard Specification for Continuity of Care Record (CCR), ASTM E2369 – 12, *ASTM International*, 2012.
- [4] A.Borodin, A.Pogorelov, Y.Zavyalova, The cross-platform application for arrhythmia detection // *Proc. of the 12th Conference of Finnish-Russian University Cooperation in Telecommunications Program*. Oulu, Finland, 5-9 Nov. 2012. Spb.: SUAI, 2012, pp. 26-30.
- [5] Y.Apanasik, I. Shabalina, L. Kuznetsova, V. Kuznetsov, Decision support information system for hardly diagnosing diseases // *Proc. of the 12th Conference of Finnish-Russian University Cooperation in Telecommunications Program*. Oulu, Finland, 5-9 Nov. 2012. Spb.: SUAI, 2012, pp. 3-12.
- [6] I.Shabalina, V. Chernetsky, Multimetods algorithm for finding the most reliable diagnoses of HS // *Transactions of PetrSU*. Series of Applied Mathematics and Computer Science, PetrSU, Petrozavodsk, 2001.
- [7] P.D. Hoff, *A First Course in Bayesian Statistical Methods*.Springer, 2009.
- [8] D. G. T. Denison, C.C. Holmes *Bayesian Methods for Nonlinear Classification and Regression*. Wiley. 2002.
- [9] Entropy Measures, *Maximum Entropy Principle and Emerging Applications*. Prof. Karmeshu edited. Springer. 2003.
- [10] W. Dupont *Statistical Modeling for Biomedical Researchers*. Cambridge University Press. 2002.